Drainage: A Unifying Framework for Addressing Class Uncertainty

(SUPPLEMENTARY MATERIAL)

Supplementary Note A. Details of the Toy Experiment

In this note, we present details of the experiments presented in the main paper in Fig. 2 (panels B and C).

2d Toy Experiment We build a dataset where data points are placed on a ring centered at the origin. Each data point's radius is drawn from a Gaussian distribution and its angle is class-dependent. Points of the first class (depicted in brown in Fig. 2) are relabeled with probability 0.5 to the second class (shown in blue). To classify the data, we consider a simple linear classification model $W\Phi(x)$. The function $\Phi : \mathbb{R}^2 \to \mathbb{R}^{|\mathcal{Q}|}$ is a fixed radial basis expansion defined as: $\Phi(x) = (\exp(-\gamma\|x - q\|^2))_{q\in\mathcal{Q}}$ where the set $\mathcal{Q}$ defines 400 basis centers, arranged on a 20×20 mesh grid spanning the input domain $[-5, 5]^2$, and where we set $\gamma = 0.25$. The parameter W is a matrix of size $(C + 1) \times 400$, learned from the data, and mapping the 400-dimensional radially expanded inputs to the logits for the C classes and the drainage node. This one-layer setting makes the training objective straightforward to optimize, and produces solutions that do not depend on the training parameters. We train the model until convergence and regularize the model by setting the weight decay parameter to 0.1.

MNIST Toy Experiment We employ the network used for our Open-Set Recognition task described in Section 6. The MNIST dataset is split into training and test sets. We randomly relabel classes 7, 8, and 9 to labels 0–6 with uniform probability, simulating missing class instances in the training data. The test set labels are kept unchanged. We evaluate the trained model under two conditions: with closed drainage $\max_i p_i$ and without open drainage $\max_{i\notin\{d\}p_i}$. Reported confusion matrices are presented in Figure 2.

Supplementary Note B. Proofs of Propositions 1 and 2

Here, we give detailed proofs for the first two propositions stated in the main paper. Let us first express the drainage loss in Eq. (4) as the following composition of functions:

$$\ell(u, v) = \log\big(1 + \alpha u + \beta v\big), \qquad u(p) = \frac{p_d + p_{\mathcal{J}}}{p_t}, \qquad v(p) = \frac{p_{\mathcal{J}}}{p_d}. \tag{8}$$

The first function monotonically increasing with u and v. The remaining two functions will be reparameterized by a real value s to model reallocation across the different probability terms. To model the reallocation from $p_{\mathcal{J}}$ to p_t, as considered in Proposition 1, we reparameterize the probability terms as:

$$\begin{pmatrix} p_t \\ p_d \\ p_{\mathcal{J}} \end{pmatrix} = \begin{pmatrix} s \\ p_d \\ 1 - p_d - s \end{pmatrix} \tag{9}$$

with $0 \leq s \leq 1 - p_d$, and p_d treated as constant. This gives us the expressions

$$u(s) = \frac{1 - s}{s} \tag{10}$$

$$v(s) = \frac{1 - p_d - s}{p_d} \tag{11}$$

which are both decreasing functions of s. Application of composition rules for monotonic functions leads to the observation that $\ell(s)$ is monotonically decreasing, in other words, the proposed drainage loss decreases when reallocating probability

from non-target classes to the target class. We consider now the reallocation from $p_{\mathcal{J}}$ to p_d studied in Proposition 2, and consider for that purpose the reparameterization

$$\begin{pmatrix} p_t \\ p_d \\ p_{\mathcal{J}} \end{pmatrix} = \begin{pmatrix} p_t \\ s \\ 1 - p_t - s \end{pmatrix} \tag{12}$$

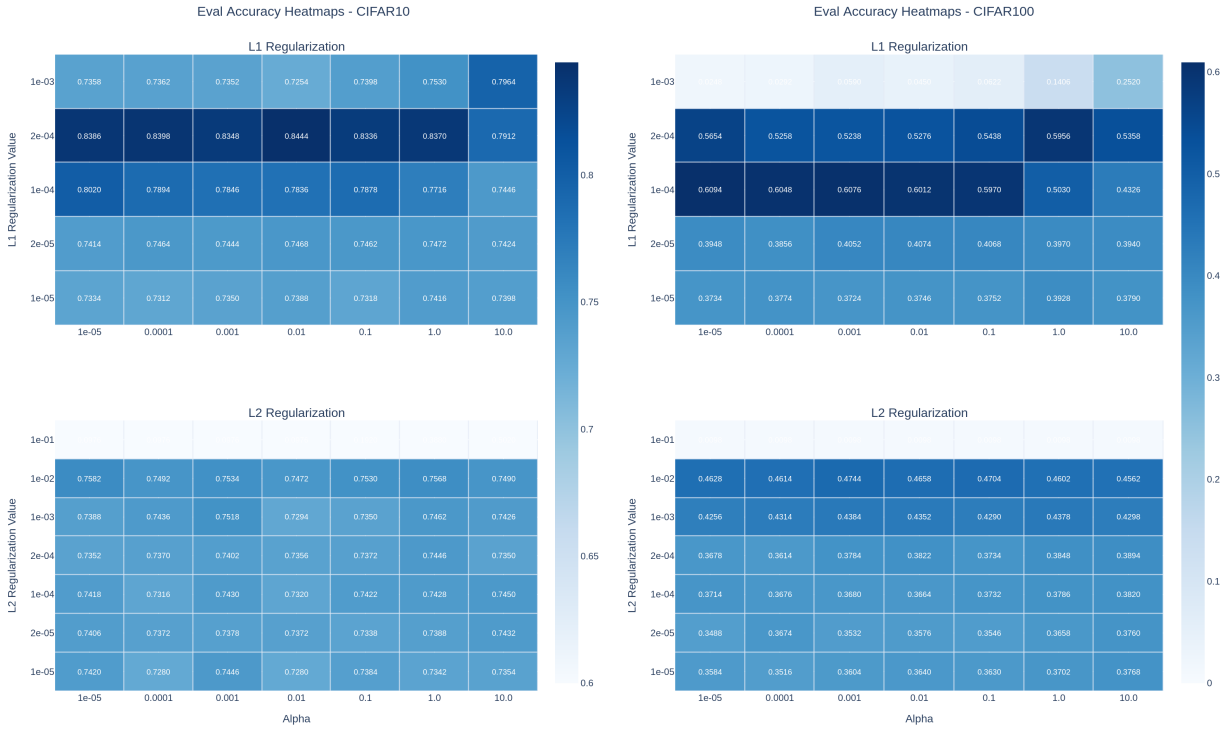
with $0 \leq s \leq p_t$, and p_t treated as constant. This gives us the expressions

$$u(s) = \alpha \cdot \frac{1 - p_t}{p_t} \tag{13}$$

$$v(s) = \beta \cdot \frac{1 - p_t - s}{s} \tag{14}$$

where $u(s)$ is constant and $v(s)$ is monotonically decreasing. Here again, applying the composition rules for monotonic functions results in the observation that $\ell(s)$ is a monotonically decreasing function. In other words, any reallocation from the non-target classes to the drainage results in a decrease of the loss.

Supplementary Figure 1. Effect of changing α and β in the MNIST toy example on the percentage of samples predicted as drainage per class. MNIST toy example depicted in Figure 2. In-distribution classes are in blue and Out-of-distribution in red.



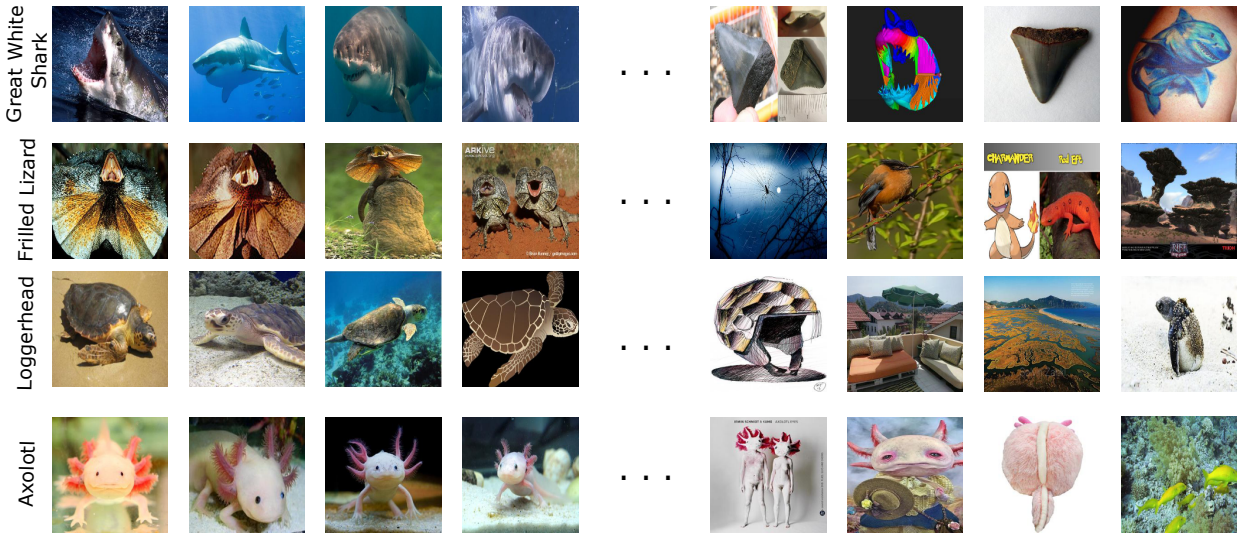
Supplementary Figure 2. Ablation study of the drainage loss on CIFAR-10 (left) and CIFAR-100 (right). We evaluate the effect of different regularization strengths, testing L1 and L2 coefficients of 1×10^{-4}, 2×10^{-4}, 1×10^{-5} and 2×10^{-5}. We grid-search for the parameter α while setting $\beta = \alpha^{-1}$.

Supplementary Table 1. Parameters used per method on each dataset based on the results presented in the main manuscript. Due to the limitation, we have not fine-tuned SCE, APL and ALF losses on Clothing1M dataset. Aside from that, for Webvision and Clothing1M, the best parameters for all methods were obtained from [31].

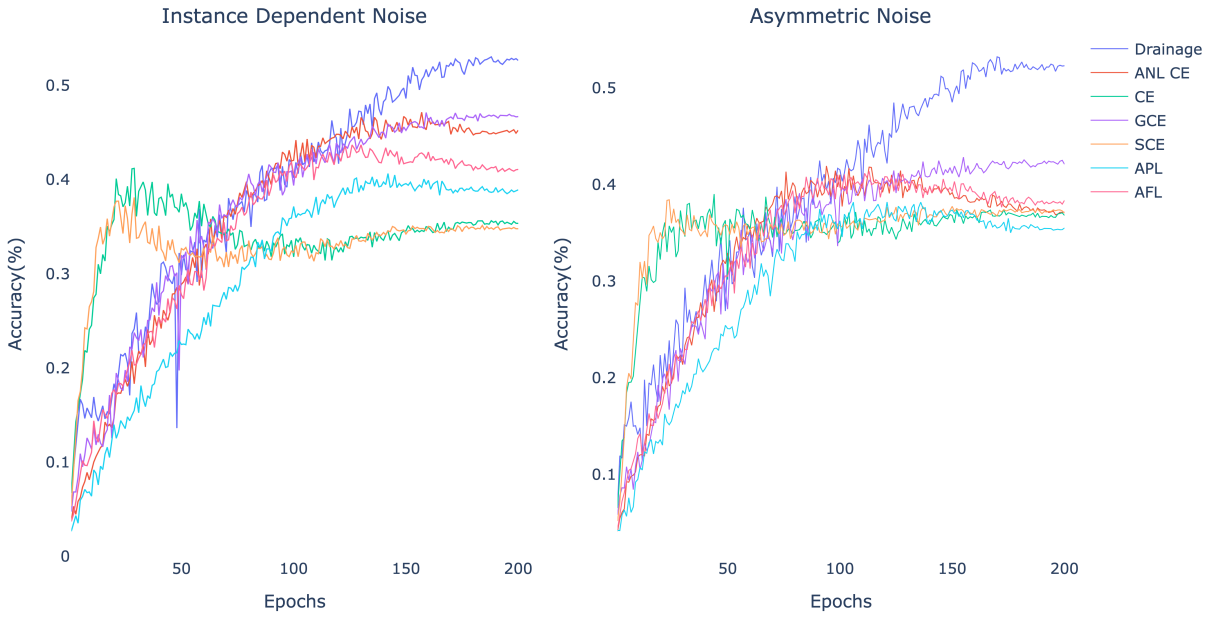
Method	CIFAR-10	CIFAR-100	Webvision	Clothing1M
CE(δ)	L1: 1×10^{-4}	L1: 1×10^{-5}	L2: 3×10^{-5}	L1: 1×10^{-3}
SCE (α, β, δ)	(0.1, 1.0, L1: 1×10^{-4})	(6.0, 0.1, L1: 1×10^{-5})	(10.0, 1.0, L2: 3×10^{-5})	(5.0, 1.0, L1: 1×10^{-3})
GCE (q, δ)	(0.7, L1: 1×10^{-4})	(0.7, L1: 1×10^{-5})	(0.7, L2: 3×10^{-5})	(0.7, L1: 1×10^{-3})
AFL ($\alpha, \beta, a, q, \delta$)	(1.0, 4.0, 6.0, 1.5, L2: 1×10^{-4})	(10.0, 0.1, 1.8, 3.0, L2: 1×10^{-5})	(50, 0.1, 2.5, 3.0, L2: 3×10^{-5})	(50, 0.1, 2.5, 3.0, L1: 1×10^{-3})
APL (α, β, δ)	(1.0, 1.0, L2: 1×10^{-4})	(10, 0.1, L2: 1×10^{-5})	(50, 0.1, L2: 3×10^{-5})	(50, 0.1, L1: 1×10^{-3})
ANL-CE (α, β, δ)	(5.0, 5.0, L1: 5×10^{-5})	(10.0, 1.0, L1: 5×10^{-7})	(20.0, 1.0, L1: 5×10^{-6})	(5.0, 1.0, L1: 1×10^{-3})
Drainage (α, β, δ)	(1.0, 1.0, L1: 2×10^{-4})	(0.1, 10.0, L1: 1×10^{-4})	(0.1, 10.0, L1: 1×10^{-5})	(0.01, 100, L1: 1×10^{-3})

Supplementary Table 2. Ablation study, where we test the same methods and datasets as in Table 1 of the main paper, but where we change the regularization scheme. Specifically, CE, GCE, SCE, AFL and APL are computed here using L2 regularization where as ANL CE and Drainage are computed using L1 regularization. For each method, we observe a decrease in performance compared to the results reported in the main paper.

	CE	GCE	SCE	AFL	APL	ANL CE	Drainage	max. err.
CIFAR-10								
noise = 0.0	90.29	89.37	91.32	91.17	91.12	91.76	91.30	±*0.04*
asymmetric:								
noise = 0.2	83.09	85.46	86.26	88.59	88.45	89.34	88.54	±*0.25*
noise = 0.3	78.51	79.83	80.53	85.48	84.89	85.13	87.56	±*0.32*
noise = 0.4	73.38	72.78	73.59	78.19	77.48	78.07	84.66	±*0.39*
noise = 0.45	70.79	69.91	70.65	71.74	72.00	72.53	80.23	±*1.41*
instance dependent:								
noise = 0.2	75.31	85.58	83.58	88.09	88.26	88.78	87.92	±*0.34*
noise = 0.4	57.07	63.98	63.86	76.29	76.57	76.54	81.95	±*0.90*
noise = 0.5	46.48	50.87	50.66	48.86	53.62	58.41	64.22	± *0.60*
CIFAR10-N	61.48	74.98	73.78	80.07	79.80	80.54	79.85	±*0.13*
CIFAR-100								
noise = 0.0	70.11	62.10	70.36	68.94	68.19	70.15	73.31	±*1.29*
asymmetric:								
noise = 0.2	59.02	64.42	58.58	63.72	62.83	66.09	70.04	±*0.87*
noise = 0.3	50.48	62.16	50.27	56.63	55.52	59.85	67.93	±*1.60*
noise = 0.4	41.79	54.44	41.48	44.23	42.60	46.01	61.55	±*1.67*
noise = 0.45	37.34	44.07	36.80	37.39	35.15	37.60	52.69	±*3.10*
instance dependent:								
noise = 0.2	57.99	62.34	57.22	64.60	63.61	66.20	67.16	±*1.23*
noise = 0.4	43.19	60.69	43.31	52.46	50.28	56.55	60.63	±*1.20*
noise = 0.5	35.31	47.21	34.61	42.11	40.52	45.67	53.01	±*1.59*
CIFAR100-N	49.57	51.31	48.64	55.29	54.62	56.62	57.77	±*0.13*



Supplementary Figure 3. Same visualization as in Fig. 3 of the main paper, but where we consider this time the training samples ordered from lowest to highest drainage probability p_d. The training data has a higher fraction of noisy labels, and they predictably end up in the rightmost columns, i.e. with high drainage probability p_d.



Supplementary Figure 4. Test accuracies of loss functions on CIFAR-100 under instance dependent noise (50%) and asymmetric noise (45%). Paramters of loss functions are as based on Table 1.

Supplementary Table 3. Results on Open Set Recognition: Closed Set Accuracy

Loss function	SVHN	CIFAR10
CE	94.6±0.9	80.6±8.9
Drainage (**ours**)	**95.5±0.6**	**84.6±3.3**

Supplementary Table 4. Performance (top-1 acc %) of CE and Drainage on ImageNet-1k. Reported results are averaged over three runs. We conducted experiments on ImageNet-1k using pretrained Vi-LSTM-tiny [1] and DeiT [24] backbones, reinitializing their classification heads and fine-tuning the models end-to-end with Drainage (default parameters: α=β=1). Incorporating Drainage improved the top-1 accuracy for both architectures.

Model	CE	Drainage (**ours**)
ViL-T	77.6±0.52	**78.1±0.61**
DeiT-T	73.8±0.17	**74.3±0.23**